Chapter 1

Frederick Boehm

February 24, 2019

Todo list

List of Figures

List of Tables

Identification of genes that affect measurable traits has a long and successful history in model organisms. Complex traits include clinical measurements such as systolic blood pressure and body weight as well as newly attainable biomolecular traits like gene expression levels, protein concentrations, and lipid levels. Understanding the genetic underpinnings of such traits may inform biology, medicine, and public health.

We begin by discussing quantitative trait locus (QTL) mapping in two-parent crosses. We then discuss pleiotropy testing in two-parent crosses before considering newly developed multiparental populations. We discuss the design for one multiparental population, the Diversity Outbred mice. We then justify the need for a pleiotropy test for use in multiparental populations.

QTL mapping in two-parent crosses

QTL mapping is a systematic, statistical approach to identifying genetic loci where genetic variation affects phenotypic variation in a measured trait. We use the term “locus” to refer to a small, contiguous genomic region, often several megabases in length. Standard inputs are genome-wide marker genotypes for a collection of study subjects and a set of trait measurements on the same subjects.

Two major statistical challenges are what Karl W Broman and Sen (2009) call the “missing data” problem and the “model selection” problem. The “missing data” problem arises in QTL studies because genotypes are obtained at only select markers. In this sense, genotypes at positions between markers are “missing”. A flexible framework for resolving the “missing data” problem involves a hidden Markov model (Karl W Broman and Sen, 2009). In this hidden Markov model, marker genotypes are observed random variables, while genotypes at intervening bases are unobserved random variables (Karl W Broman and Sen, 2009; Karl W. Broman, 2006). Two sets of recursive equations, termed “forward” and “backward” equations, enable efficient calculation of genotype probabilities (conditional on the observed marker genotypes) (**baum1970maximization**).

In anticipation that multivariate mapping of correlated traits would enhance statistical power to detect QTL and would improve precision of QTL positions, both Jiang and Zeng (1995) and **korol1995correlated** developed multivariate interval mapping procedures.

In the context of multivariate QTL mapping, Jiang and Zeng (1995) proposed and developed a statistical hypothesis test for distinguishing a single pleiotropic locus that affects multiple traits from two (or more) separate QTL. Their test applies to the setting in which multiple traits map to a single genomic region (Jiang and Zeng, 1995). That region, then, can be interrogated further by asking whether it harbors a single pleiotropic QTL or multiple separate QTL. For simplicity, we focus on the case where two traits both map to a single genomic region, and we wish to determine whether the two traits share a single QTL or have two separate QTL.

Diversity Outbred Mice

Geneticists sought a mammalian gene mapping resource that could be used for study of a wide variety of quantitative traits. The magnitude of such an undertaking required a collaborative, community-supported approach (Koning and McIntyre, 2014). Scientists conceived of the Diversity Outbred (DO) mouse population as such a high-resolution gene mapping resource. They elected to seed the population with partially inbred progenitors of the Collaborative Cross mouse population. The Collaborative Cross mouse mating design involved starting with mice from eight inbred lines: A/J, C57BL/6J, 129S1/SvImJ, NOD/ShiLtJ, NZO/HILtJ, CAST/EiJ, PWK/PhJ, WSB/EiJ.

The designers used a “funnel” mating scheme to generate mice with DNA from all eight founder lines over the course of 3 generations. For example, in one funnel, mating pairs are: A x B, C x D, E x F, and G x H in the first generation. AB offspring would then mate with CD offspring and EF offspring would mate with GH mice. Finally, the ABCD mice would mate with the EFGH mice to create a generation of mice that contain genetic material from all eight inbred founder lines. Subsequent generations of inbreeding resulted in multiple inbred lines for the Collaborative Cross.

make this figure

The designers of the Diversity Outbred population started with

get number of mice and from which generations

Due to the breeding scheme, each Diversity Outbred mouse is a highly heterozygous mosaic of founders' DNA (Figure). With each generation, the Diversity Outbred mice accumulate meiotic recombinations. This is because each mouse inherits DNA from its parents, and each meiosis provides opportunities for recombinations.

add figure with mosaics over multiple generations here

need to explain this better. How does Andrew explain it in his thesis?

The cumulative effect of recombinations over many generations of outbreeding is to create mice whose DNA mosaic contains smaller and smaller contiguous pieces from each founder line.

The Diversity Outbred mice enable high-resolution QTL mapping (Gatti et al., 2014). One factor that contributes to high-resolution mapping is the ability to infer the founder line from which each marker's DNA arose.

QTL mapping in DO mice

QTL mapping in Diversity Outbred mice, like that in mice from two-parent crosses, is a multi-step procedure: 1. data acquisition, 2. inference of missing genotypes, and 3. modeling phenotypes as a function of genotypes. Data acquisition involves measurement of phenotypes and, at specified genetic markers, termed “single nucleotide polymorphism” or SNP markers, measurement of two-allele genotypes. Often, the SNP marker genotypes are obtained by use of a microarray, such as the GigaMUGA SNP microarray (Morgan et al., 2015).

The next step, missing genotypes inference, is needed because of the “missing data problem” (Karl W Broman and Sen, 2009). It takes as input the two-allele genotypes at the measured SNP markers. An expectation maximization algorithm (Dempster, Laird, and Rubin, 1977) for a hidden Markov model, developed by Karl W Broman (2012b) and Karl W Broman (2012a) and implemented in the `qtl2` R package (K. W. Broman et al., 2019), outputs 36-state genotype probabilities for all nuclear autosomal markers and pseudomarkers. Pseudomarkers, as we use the term, are arbitrary nucleotide bases at which the researcher wants 36-state genotype probabilities. Finally, we collapse the 36-state genotype probabilities to eight founder allele dosages at each marker and for every mouse before performing linear regression to identify QTL for every univariate trait.

Testing pleiotropy in two-parent crosses (Jiang & Zeng

Jiang and Zeng (1995) developed a suite of methods for analyzing multivariate phenotypes in two-parent crosses. Among the novel methods in their article is a test of pleiotropy vs separate QTL. They explain that such a test is useful when two traits map to a single genomic region. The question then arises "do the two traits associate with the same locus, or do they associate with separate loci"?

If both traits associate with the same locus, then that locus is called "pleiotropic". If the two traits associate with separate loci, then we say that they have separate QTL (one for each trait). While "pleiotropy" has multiple related usages in the genetics literature, we use "pleiotropy" here to refer to the situation in which one genetic locus affects two or more traits.

In the development of their test of pleiotropy v separate QTL, Jiang and Zeng model the multivariate phenotype as the sum of a linear function of the genotypes and a random error term. They assume that the phenotypes matrix is related to the genotypes data (for a single marker) through the following equation:

$$Y = XB + E \tag{1}$$

Where Y is a n by 2 matrix of phenotype values (with each row being one subject and each column being one trait), X is a n by 1 matrix of genotypes and B is a 1 by 2 matrix of allele effects. The random error is assumed to follow a normal distribution with mean zero and a positive variance. Jiang and Zeng (1995) also incorporate additional terms to model dominant effects and to control for residual genetic variation. We forego inclusion of dominant terms and additional markers for controlling residual genetic variation.

Jiang and Zeng (1995) use a mixture model to relate genotypes to phenotypes. They observe that the there are 9 possible ordered pairs of genotypes for two markers in a F2 population. They thus approach the problem as one with a 9-state mixture model. They provide the equations needed for an expectation-conditional maximization (ECM) algorithm for fitting the mixture model. They use a chi-square distribution with one degree of freedom as the null distribution of the test statistic.

Further investigations by Jiang and Zeng (1995) with simulations revealed reasonable behavior of the hypothesis test under the specified conditions. They learned that dense marker coverage tends to aid in distinguishing pleiotropy from separate QTL.

look back at JZ 1995's simulations. How did they examine power and type I error rates? Do they examine univariate LOD strength when considering power? What about interlocus distance?

Finding values that minimize the sum of squared residuals - THINK ABOUT WHAT A RESIDUAL IS HERE... a matrix? - They use a computational algorithm called "expectation - conditional maximization" to find these values. ECM is an iterative algorithm. The user inputs starting values for the parameters and the algorithm changes them one at a time until the change in log likelihood between successive iterations is sufficiently small. The algorithm is then said to have "converged". The final iteration's parameter values are treated as those that maximize the likelihood.

Jiang & Zeng (1995), in the context of their multivariate QTL mapping methods, describe a test of pleiotropy vs separate QTL. They formulate their question as a statistical hypothesis test in which the null hypothesis is pleiotropy while the the alternative hypothesis is separate QTL. In other words, they consider the case in which the two traits map to the same locus (ie, pleiotropy) as a special case of the more general setting (separate QTL) in which the two traits may or may not map to the same locus. In statistical terminology, one would say that the parameter space is restricted under the null hypothesis.

Jiang and Zeng then calculate the likelihoods of the two models (that under the hypothesis of pleiotropy and that under the hypothesis of separate QTL) at each ordered pair of putative loci in the genomic region of interest. The likelihood ratio test statistic is the logarithm of the ratio of the maximum of the likelihoods under pleiotropy to the maximum of the likelihoods under the separate QTL hypothesis.

Jiang and Zeng determined p-values for their test statistics by comparing them to a theoretical null distribution. In this specific case, Jiang and Zeng reasoned that the null distribution of their test statistic is a chi-squared distribution with 1 degree of freedom. Their justification for this is based on asymptotic statistics theory.

Schadt et al (2005) extended the methods of Jiang and Zeng (1995).

References

Broman, K. W. et al. (2019). "R/qtl2: Software for mapping quantitative trait loci with high-dimensional data and multi-parent populations". In: *Genetics, to appear.*

Broman, Karl W (2012a). "Genotype probabilities at intermediate generations in the construction of recombinant inbred lines". In: *Genetics* 190.2, pp. 403–412.

– (2012b). "Haplotype probabilities in advanced intercross populations". In: *G3: Genes, Genomes, Genetics* 2.2, pp. 199–202.

Broman, Karl W. (2006). *Use of hidden Markov models for QTL mapping.* Tech. rep. 125. Johns Hopkins University Department of Biostatistics. URL: http://biostats.bepress.com/jhubiostat/paper125.

Broman, Karl W and Saunak Sen (2009). *A Guide to QTL Mapping with R/qtl.* Vol. 46. Springer.

Dempster, Arthur P, Nan M Laird, and Donald B Rubin (1977). "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38.

Gatti, Daniel M et al. (2014). "Quantitative trait locus mapping methods for diversity outbred mice". In: *G3: Genes, Genomes, Genetics* 4.9, pp. 1623–1633.

Jiang, Changjian and Zhao-Bang Zeng (1995). "Multiple trait analysis of genetic mapping for quantitative trait loci." In: *Genetics* 140.3, pp. 1111–1127.

Koning, Dirk-Jan de and Lauren M McIntyre (2014). "GENETICS and G3: Community-Driven Science, Community-Driven Journals". In: *Genetics* 198.1, pp. 1–2.

Morgan, Andrew P et al. (2015). "The mouse universal genotyping array: from substrains to subspecies". In: *G3: Genes| Genomes| Genetics*, g3–115.